Please cite as:

Onghena, P. (2018). Randomization and the randomization test: Two sides of the same coin. In V. Berger (Ed.), *Randomization, masking, and allocation concealment* (pp. 185-207). Boca Raton/FL: Chapman & Hall/CRC Press.

ISBN-13: 978-1-138-03364-1

Randomization and the randomization test: Two sides of the same coin

Patrick Onghena

KU Leuven – University of Leuven, Belgium

Abstract

This chapter provides a framework for conceptualizing randomization in clinical trials and for linking randomization to a statistical test. The aim of this chapter is to demonstrate how randomization works, how it does not work, and why a discussion of randomization without reference to a randomization test is incomplete. Randomization works because, on average, the differences between the treatment averages of any potentially confounding variable are zero. Randomization does not work in the sense that it does not contain any magic equating potion that spirits away all biases for any particular clinical trial. Furthermore, it should be taken into account that randomization is closely linked to statistical inference by the concept of a randomization test. This test is formally defined and its main features are highlighted: The randomization test is valid and powerful by construction, it can be used with any design and any test statistic, without random sampling and without assuming specific population distributions, and it results in a frequentist, conditional, and causal inference. The randomization test derives its statistical validity by virtue of the actual randomization and conversely a randomized design is complemented with a calculation of the randomization test *p*-value because it provides a quantification of the probability of the outcome (or a more extreme one) if the null hypothesis is true.

The idea of randomization for eliminating bias in empirical research and the proposal for using the corresponding technique of a randomization test are relatively recent methodological developments, taking into account the long history of the scientific enterprise (David, 1995; Folks, 1984; Hacking, 1988; Salsburg, 2001; Stigler, 1978, 1986). In the 19th century, randomization was introduced in psychophysical investigations to keep participants unaware of the experimental manipulations and thus avoiding systematic distortions due to participants' knowledge or their willingness to act (or not) according to the researcher's hypotheses (see e.g., Peirce & Jastrow, 1885), but it was only in the first half of the previous century that Fisher (1925, 1926, 1935) convincingly started promoting randomization as an essential ingredient of any proper experimental research plan and as a way to validate statistical tests.

In medical research, Sir Bradley Hill has been given credit for setting the randomized controlled trial as the gold standard for comparing treatment effects, with the Streptomycin study in 1948 as the exemplar (Cochrane, 1972; Matthews, this volume). Nowadays, randomization belongs to the canon of valid empirical research in evidence-based medicine (Begg et al., 1996; Higgins & Green, 2011; Jadad & Enkin, 2007). However, from its inception, and continued throughout the second half of the previous century, (mainly Bayesian) statisticians have seriously questioned the use of randomization, its methodological benefits, and its implications for statistical analysis (Basu, 1975, 1980; Kadane & Seidenfeld, 1990; Lindley, 1982; Savage, 1962; Stone, 1969), with doubts remaining up to this very day (see Saint-Mont, 2015, for an overview).

Because randomization is only recently adopted in the realm of science and because experts in methodology and statistics are still debating its virtues and vices, it is no surprise that there remain several misunderstandings, resistance, and suboptimal surrogates to the idea of randomization in scientific research (Berger & Bears, 2003; this volume). In this chapter we want to contribute to the ongoing debate by providing a framework for conceptualizing randomization and by linking randomization to a statistical test. Our aim is to show how randomization works, how it does not work, and why a discussion of randomization without reference to a randomization test is incomplete.

RANDOMIZATION

What is randomization and how does it work? Perhaps more importantly: how does it *not* work?

Randomization in clinical trials

Given the topic of this book, the presentation will be restricted to randomization in clinical trials. So the experimental units are patients, and the "randomization" refers to the random assignment of patients to the treatments that are compared in the clinical trial (Jadad & Enkin, 2007). If the number of patients in the trial is denoted by N, the number of treatments is denoted by J, and the number of patients assigned to respectively treatment $T_1, T_2, \dots T_j, \dots T_J$ is denoted by $n_1, n_2, \dots n_j, \dots n_J$, then the total number of possible distinct assignments of patients to treatments, denoted by K, is given by the multinomial coefficient:

$$K = \frac{N!}{\prod_{j=1}^{J} n_j!} \tag{1}$$

If there are only two treatments, $J = 2$, for example in a double-blind clinical trial of a medical agent versus a placebo control, the total number of possible assignments simplifies to a binomial coefficient:

$$K = \binom{N}{n_1} \tag{2}$$

Randomly assigning, or what Fisher (1935, p. 51) called "the physical act of randomisation", can be conceptualized as taking a simple random sample of size one out of the population of *K* possible assignments, with the population of possible assignments generated according to the design specifications (Finch, 1986; Kempthorne, 1986). Equivalently, if the *N* available patients are considered as the population, the random assignment can be conceptualized as taking a sample without replacement of size *N* and respecting the order of the sampled patients, the order of the treatments, and the number of patients assigned to the treatments (Efron & Tibshirani, 1993; Good, 2005).

Why randomize?

There are many reasons why researchers include randomization in their empirical studies, and reasons may vary between disciplines (Greenberg, 1951; Kempthorne, 1977; Rubin, 1978; Senn, 1994). In double-blind clinical trials, randomization is included to safeguard the double-blindness, to be consistent with the principle of equipoise, and to take into account ethical considerations with respect to informed consent and with respect to a fair allocation of decent and affordable health care services (Begg, 2015; Freedman, 1987; Matthews, this volume.).

From a general statistical perspective, there are two closely related reasons given in the literature: the elimination of bias and the justification of a statistical test (Cochran & Cox, 1950; Cox & Hinkley, 1974; Kempthorne, 1952, 1955; Kempthorne & Folks, 1971; Rosenberger & Lachin, 2015). In this section we will focus on the elimination of bias. In the next section, we will demonstrate how randomization may justify a statistical test.

The property of bias elimination can be positively reformulated as the equating property of randomization. Randomization is used in clinical trials to equate the groups that are assigned to the different treatments in a statistical way. As Rubin (2008) put it:

> Another reason why randomized experiments are so appealing (…), is that they achieve, in expectation, "balance" on all pre-treatment-assignment variables (i.e., covariates), both measured and unmeasured. Balance here means that within well-defined subgroups of treatment and control units, the distributions of covariates differ only randomly between the treatment and control units. (p. 809)

By using randomization, treatments become comparable because the patients are exchangeable. Because systematic differences between the patients can be ignored, the differences between the treatment outcomes can be attributed to the differences between the treatments and not to any differences between the patients accidentally assigned to the treatments.

And here the confusion begins. Researchers might be tempted to use randomization as a magical device that automatically washes out any difference between patient groups in a single clinical trial (see Saint-Mont, 2015, for some examples). However, the equating property only holds in a statistical way. In the Rubin (2008) quote it is important to notice the phrase “in expectation” (and to interpret this expectation in a strict statistical way) and that “the distributions of covariates” still differ “between the treatment and control units”, be it “only randomly”.

Clarification of the equating property of randomization

What do this “statistical way” and “expectation” mean? They mean that *on the average* the differences between the treatment averages of any potentially confounding variable are zero. Suppose that we use a completely randomized design with the number of possible assignments of *N* patients to *J* treatments equal to *K*, as given in Equation (1). Take any confounding variable *Z*, with z_{ij} as the measurement of that variable of patient *i* in treatment *j.* Randomly assigning patients to treatments implies randomly assigning the measurements of the confounding variable to the treatments, so let z_{ijk} be the measurement of patient *i* in treatment *j* for assignment *k*.

Consider a particular assignment of the *N* measurements to the *J* treatments. Then consider all *K*−1 rearrangements of these measurements to the treatments and organize them in a large *N* × *K* matrix. In such a matrix the total sum of all measurements is equal to *K* times the sum of the observed measurements (or any other rearrangement of the measurements) because the sum is a constant for each assignment:

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ijk} = K\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ij} \tag{3}$$

The total sum of all measurements in the matrix is also equal to $\frac{N}{n_j}$ times the sum of all measurements within a particular treatment in the matrix, z_{ik}, computed over all possible assignments, because all measurements occur an equal number of times within a treatment:

$$\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ijk} = \frac{N}{n_j}\sum_{k=1}^{K}\sum_{i=1}^{n_j} z_{ik} \tag{4}$$

Working back from Equation (4) and substituting for Equation (3) gives:

$$\sum_{k=1}^{K}\sum_{i=1}^{n_j} z_{ik} = \frac{n_j}{N}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ijk} = \frac{n_j K}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ij} \tag{5}$$

Consequently, if you randomly select one assignment out of the total number of possible assignments, then the expected value of the average of the measurements within each treatment is equal to the average of the *N* measurements:

$$E\left(\frac{\sum_{i=1}^{n_j} z_i}{n_j}\right) = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_j} z_{ik}}{Kn_j} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j} z_{ij}}{N} \tag{6}$$

This implies that, on average, all treatment averages are identical and in this sense the groups in a completely randomized design are equated. In other words, the bias due to potentially confounding variables is, "in expectation", eliminated.

An example

A numerical example may clarify this obvious, but nevertheless easy to misunderstand, property. Suppose that you are testing a new diet against a control and that four patients are available. You use a completely randomized design with $N = 4$, $J = 2$, $n_1 = 2$, $n_2 = 2$, and therefore $K = 6$. The weight of the patients before the trial is evidently a confounding variable; suppose that you try to control this confounding variable by randomization. Commonly, the numerical values of the confounding variables are unknown, but suppose in this case, for the purpose of demonstration, we have put the patients on a scale before the trial is conducted and registered the weights for the four patients to be respectively 62 kg, 75 kg, 76 kg, and 127 kg. The matrix of all possible assignments and the marginal averages are given in Table 1. Table 2 shows the averages for each treatment and the differences between the averages for each assignment.

_____

Insert Table 1 here

_____

_____

Insert Table 2 here

_____

It turns out that the average (i.e., the expected value) of the averages of the measurements for treatment as well as for control is equal to 85. This is exactly the average of the four numbers 62, 75, 76, and 127 (and also the overall average of all numbers in the matrix). Notice also in Table 2 that any specific assignment violates the assumed magical equating property: the deviation between the treatment average and the control average ranges from 33 to −33 kg. It is only on average that the deviation is zero. So it is only on average that the equating property of randomization works. It is not guaranteed to work for any single assignment.

Concluding remarks on the equating property of randomization

It is sometimes argued that the magical equating property for single assignments *will* work if a larger number of patients is recruited. However, this is a red herring. Although the deviations for one particular assignment and one particular confounding variable might decrease by increasing the number of patients, the sheer number of potential known and unknown confounding variables and their interactions is close to infinity. So the probability that there will show up an imbalance for at least one confounding variable is close to one.

This clarification of the equating property of randomization and the small numerical example may seem trivial and nonrealistic but at the same time they show the deeper and limited benefits of randomization. Too many handbooks in statistics give only cursory attention to the "physical act" and in this way turn randomization into a design ritual or, even worse, into a reporting ritual that has no bearing on reality. By contrast, our elaboration and Equation (6) show that the equating property

holds for the very broad class of completely randomized designs, so even for designs that are, what is called, “unbalanced” (i.e., designs with an unequal number of patients in the treatments). So even unbalanced designs “achieve, in expectation, ‘balance’ on all pre-treatment-assignment variables (i.e., covariates), both measured and unmeasured”. In addition, the example illustrates that the property does not need nicely behaving Gaussian or symmetrically distributed confounding variables or asymptotic derivations. The property already works with only two patients in each of two treatments, even with a blatantly skewed confounding variable.

So as a conclusion, the equating property is one of the main reasons that randomization deserves a place in the clinical trials design toolbox. However, we must not get carried away by statistical enthusiasm: This equating property does not guarantee optimal balance and comparability in any particular assignment. Furthermore, if a researcher has prior knowledge about the patients or about the effectiveness of the treatments, then other ways of equating may be used, such as matching or blocking (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002) or using adaptive allocation rules (Berry, Carlin, Lee, & Müller, 2010; Hu & Rosenberger, 2006). Just to be sure, randomization was never proposed as a panacea solution for all evils. Randomization is at its best in conjunction with other methodological maneuvers to gain control and complemented by a statistical test based on the randomization as it was actually carried out.

THE RANDOMIZATION TEST

> In these discussions it seems to have escaped recognition that the physical act of randomisation, which, as has been shown, is necessary for the validity of any test of significance, affords the means, in respect of any particular body of data, of examining the wider hypothesis in which no normality of distribution is implied. (Fisher, 1935, p. 51)

In this quote from the groundbreaking *Design of Experiments*, Fisher (1935) is opposing the critics of Student's *t* test who emphasize that the assumption of normally distributed populations represents a serious limitation to its applicability. He continues by proposing a statistical test for data on the difference between the relative growth rates of cross- and self-fertilized corn, collected by Charles Darwin and analyzed by Francis Galton, in which no normally distributed populations are assumed. He calls this a test of “the wider hypothesis” that the two distributions are identical. This test is a precursor of the significance test “which may be applied to samples from any populations” proposed by Pitman (1937a, 1937b, 1938) and which has later been termed a “randomization test” (Berry, Johnston, & Mielke, 2014; David, 2008, Kempthorne, 1952, 1955) but for our purposes it suffices to join Fisher (1935) in emphasizing that:

- Randomization is needed for the validity of any statistical test;
- The act of randomization is physical; and
- There is no need for distributional assumptions if you want a test of the wider hypothesis that the two distributions are identical.

So an account of the purposes of randomization is incomplete without reference to the statistical test for which it was originally devised. Because the equating property of randomization only describes a “statistical” regularity, it needs to be complemented by a statistical test to minimize

erroneous scientific inferences based on the observed data; the statistical test minimizes these erroneous scientific inferences to a known (and small) degree. Furthermore, the foundation of the statistical test is an operation of randomization that is actually carried out, not just a mere thought experiment or assumption. Incidentally, this "randomization" test also relieves the researcher from assuming normally distributed populations; it is the distribution-free statistical test *par excellence*.

In this section, we briefly present the rationale of this test based on randomization and discuss its main features. For more extensive presentations and discussion, the reader is referred to Good (2005), Manly (2007), and Edgington and Onghena (2007).

Notation

We will develop the notation for a completely randomized design with a simple treatment-control comparison involving *N* available patients. However, the notation can easily be extended for completely randomized designs comparing two treatments, for completely randomized designs comparing three treatments or more, or for other designs.

Let $\boldsymbol{x}' = (x_1,\ x_2,\ \dots,\ x_i, \dots,\ x_N)$ be a design vector with $x_i = 1$ if patient *i* is assigned to Treatment and $x_i = 0$ if patient *i* is assigned to Control. The randomness of the assignment involves the random selection of one particular design vector $\boldsymbol{x'_k} = (x_{1k},\ x_{2k},\ \dots,\ x_{ik}, \dots,\ x_{Nk})$ from the set of all possible design vectors $\boldsymbol{X} = \{\boldsymbol{x_1},\ \boldsymbol{x_2},\ \dots,\ \boldsymbol{x_k}, \dots,\ \boldsymbol{x_K}\}$. The randomly selected design vector is called the observed design vector and denoted by $\boldsymbol{x_{obs}}$, and the set of all possible design vectors $\boldsymbol{X}$ is called the reference set. The cardinality of the reference set, *K*, with n_1 patients assigned to Treatment, leaving $N - n_1$ patients for Control, was given in Equation (2).

If the trial contains a comparison of $J > 2$ treatments, then $J - 1$ design vectors and additional subscripts are needed. It is also possible that the x_i values represent the "dose" of the treatment if the treatment is manipulated quantitatively. In that case only one design vector is needed, even with $J > 2$. The cardinality of the reference set for a completely randomized design with *J* treatments was given in Equation (1). The cardinality of the reference set for a randomized design in which the treatment is manipulated quantitatively is equal to *N*! (see Chapters 8 and 9 of Edgington & Onghena, 2007, for some interesting applications for designs with quantitatively manipulated treatments).

Furthermore, consider $\boldsymbol{u}' = (u_1,\ u_2,\ \dots,\ u_i, \dots,\ u_N)$, a vector of basic responses, and $\boldsymbol{y}' = (y_1,\ y_2,\ \dots,\ y_i, \dots,\ y_N)$, a vector of actually observed responses. The basic responses are the responses that the patients would give if the patients are assigned to Control (c.q. if the treatment has no effect). These basic responses are transformed to actually observed responses, taking into account treatment assignment and potential treatment effects.

The null hypothesis and several possibilities for the alternative hypothesis

The null hypothesis H_0 of the randomization test in a completely randomized design for a comparison of Treatment and Control states that there is no treatment effect:

$$H_0: \boldsymbol{y} = \boldsymbol{u} \tag{7}$$

In other words, if the null hypothesis is true, we observe the basic responses. This formula has to be slightly adapted or extended if two or more treatments are compared, to take into account that in

that case the null hypothesis is formulated in terms of "no differential effects of treatments", but this does not alter the basic reasoning and procedure.

Edgington (1986, p. 531) defined the alternative hypothesis of a randomization test as the hypothesis that at least one patient would have responded differently under one of the other treatments. This is consistent with the Fisherian approach, in which the alternative hypothesis is just the complement of the null hypothesis. In the Neyman-Pearson approach, however, one should try to consider more specific alternative hypotheses.

The simplest model for specific alternative hypotheses, following the Neyman-Pearson approach, is that the treatment has a constant additive effect (denoted by Δ) on the responses. This model is called the "unit-treatment additivity model" and it is the model that is most popular and well-studied within nonparametric statistics (see e.g., Cox & Reid, 2000; Hinkelmann & Kempthorne, 2005, 2008, 2012; Lehmann, 1975):

$$H_1: \boldsymbol{y} = \boldsymbol{u} + \Delta \boldsymbol{x_k} \tag{8}$$

For example, if the treatment adds one measurement unit to each basic response if the patient is assigned to the treatment, the null and alternative hypothesis correspond to:

$$\begin{aligned} H_0&: \Delta = 0 \\ H_1&: \Delta = 1 \end{aligned} \tag{9}$$

The unit-treatment additivity model presupposes that there is no unit-treatment interaction: Δ is a constant.

A model that allows for unit-treatment interaction is the multiplicative model. In the multiplicative model the treatment effect is proportional to the basic responses:

$$H_1: \boldsymbol{y} = \boldsymbol{u}(\mathbf{1} + \tau \boldsymbol{x_k}) \tag{10}$$

For example, if one assumes that the treatment doubles each basic response for the patient that receives the treatment, the null and alternative hypothesis correspond to:

$$\begin{aligned} H_0&: \tau = 0 \\ H_1&: \tau = 1 \end{aligned} \tag{11}$$

An infinite number of other models can be conceived of. Models that contain extreme unit-treatment interaction are models in which **Δ** (or **τ**) is a vector, with Δ_i (or τ_i) varying from $i = 1$ to N. In such a model each patient has his/her own additive or multiplicative treatment response. It goes without saying that in most cases researchers would be very hard-pressed to produce such detail in the alternative hypothesis, with all the Δ_i specified before the trial is conducted. If it were possible to produce such detail, then one could even question the need for conducting the trial in the first place.

The typical Fisherian solution (echoed by Edgington, 1986) is to set the lower bound (i.e., all Δ_i equal to zero, except for one) and to consider the complete set of alternative hypotheses in an unrestricted unit-treatment interaction model. An intermediate solution, which can be used for conditional power analysis, is to assume that the Δ_i or τ_i represent a random sample from a particular well-described

population distribution of treatment effects, such as the exponential distribution used in Keller's (2012) simulation study to investigate the power of randomization tests when *N* is small.

The test statistic

Like any other statistical test, the randomization test needs a test statistic. Unlike any other statistical test, however, the researcher is free in his choice of this statistic.

A test statistic is needed because the researcher has collected a batch of data, wants to summarize these data, and wants to know whether or not a treatment effect can be inferred from the data summary, possibly also how large this effect is, given the complete batch. This statistic can be chosen freely by the researcher in accordance to his theory about the possible treatment effect (the alternative hypothesis discussed in the previous section), but for the validity of the procedure it is of paramount importance that the choice is made independently from knowing the data. Usually this independence is assured by making the choice before the data are collected, but if appropriate masking techniques are used, then it is also possible to postpone this choice until a moment before the data are disclosed to the researcher or the data analyst (see Ferron, & Foster-Johnson, 1998; Ferron, & Jones, 2006; Ferron & Levin, 2014, for some interesting applications of data masking).

Formally, the test statistic s is a function of the design vector and the response vector: $s = s(\boldsymbol{x}, \boldsymbol{y})$. If for example, the unit-treatment additivity model of Equation (8) is assumed in a completely randomized design with n_1 patients assigned to Treatment, leaving $N - n_1$ patients for Control, then the additive treatment effect on all patients can be captured by a difference between the average response of the treatment and the average response of the control, $s = \frac{1}{n_1}\boldsymbol{x}'\boldsymbol{y} - \frac{1}{N-n_1}(\mathbf{1}' - \boldsymbol{x}')\boldsymbol{y}$.

The randomization test p-value

The final step of a statistical test involves the calculation of the *p*-value, which is the probability to obtain a test statistic as extreme as the observed test statistic or even more extreme, conditional on the truth of the null hypothesis. The only random element in the randomization model under the null hypothesis, as it was presented in the previous paragraphs, is given by the random selection of the observed design vector, $\boldsymbol{x_{obs}}$, out of the reference set, $\boldsymbol{X}$. So the only way to calculate probabilities under the null hypothesis concerns this random selection.

Define the test statistic $s_k = s(\boldsymbol{x_k}, \boldsymbol{y})$, as a function of the design vector and the response vector; the design vector is random and the response vector is fixed. Let $s_{obs} = s(\boldsymbol{x_{obs}}, \boldsymbol{y})$ be the observed value of the test statistic, also called the observed statistic for short, or put $s_{obs} = s_1$ for convenience. The other values of the test statistic $s_2, s_3, \ldots, s_k, \ldots, s_K$ are called the additional reference statistics, and the frequency distribution on the reference statistics is called the reference distribution.

The probabilities of particular subsets of test statistics can easily be calculated from this reference distribution. More specifically, if large values of the test statistic represent effectiveness, then right-tail *p*-values are obtained as:

$$p_{right} = P(s_k \geq s_{obs}) = \frac{1}{K}\sum_{k=1}^{K} I(s_k \geq s_{obs}) \qquad (12)$$

with $I(a)$ the indicator function, returning a 1 if proposition a is true and returning 0 otherwise. Remark that the number of reference statistics that are as large as the observed statistic or larger, call this η, is at least one because the observed statistic, $s_{obs} = s_1$, is also a reference statistic:

$$\begin{aligned}\eta &= Kp_{right} \\ &= \sum_{k=1}^{K} I(s_k \geq s_{obs}) \\ &= 1 + \sum_{k=2}^{K} I(s_k \geq s_1)\end{aligned} \tag{13}$$

This implies that the minimal *p*-value of the randomization test is equal to 1/*K*.

If small values represent effectiveness, then left-tail *p*-values are obtained:

$$p_{left} = P(s_k \leq s_{obs}) = \frac{1}{K}\sum_{k=1}^{K} I(s_k \leq s_{obs}) \tag{14}$$

Two-tailed *p*-values are obtained by defining a test statistic with an absolute value or with a square and applying Equation (12).

Fisher (1925, 1935) used the *p*-value as a measure of evidence against the null hypothesis: the smaller this *p*-value, the more evidence against the null hypothesis. Notice however that this *p*-value does not quantify the probability that the null hypothesis is true. It is merely a calculation of the probability to observe the data (or even more extreme data), given that the null hypothesis is true (Edgington, 1970; Wasserstein & Lazar, 2016).

In the Neyman-Pearson approach, this *p*-value is used in a decision rule to reject or accept the null hypothesis. The null hypothesis is rejected if and only if the *p*-value is not larger than the level of significance α:

$$\text{Reject } H_0 \Leftrightarrow p \leq \alpha \tag{15}$$

Following this approach, the level of significance α is set before the trial is run and is the yardstick that a researcher wants to use as the criterion for (the largest probability of) an improbable event. If a *p*-value is obtained that is equal to this criterion or even smaller, then an improbable event has happened if the null hypothesis is true. Because a researcher does not want to believe in improbable events, (s)he rejects the null hypothesis. Traditionally, values for α of 5%, 1%, or 0.1% are used.

An example

Reconsider the hypothetical example of testing a new diet against a control and now take the patient's weight registered at the end of the trial as the outcome variable. Suppose that a researcher wants to test the null hypothesis that the treatment has no effect, with 10 consecutive patients eligible for the trial, and that (s)he chooses the difference between the averages as the test statistic. The level of significance α is set at 5%.

Five patients are randomly assigned to treatment and five patients are randomly assigned to control. Applying Equation (2) gives $K = \binom{10}{5} = 252$ design vectors in the reference set, $\boldsymbol{X}$. These design vectors are:

$\boldsymbol{x'_1} = (1,1,1,1,1,0,0,0,0,0)$

$\boldsymbol{x'_2} = (1,1,1,1,0,1,0,0,0,0)$

$\boldsymbol{x'_3} = (1,1,1,1,0,0,1,0,0,0)$

and all other combinations up to

$\boldsymbol{x'_{252}} = (0,0,0,0,0,1,1,1,1,1).$

Suppose we randomly selected the first design vector $\boldsymbol{x_1}$ and that the following response vector is obtained: $\boldsymbol{y} = (51,62,75,76,77,74,89,90,92,127)$ in kilograms. The observed statistic is $s_{obs} = s_1(\boldsymbol{x_1}, \boldsymbol{y}) = \frac{1}{5}\boldsymbol{x'y} - \frac{1}{5}(\boldsymbol{1'} - \boldsymbol{x'})\boldsymbol{y} = 68.2 - 94.4 = -26.2$. All additional reference statistics are calculated and the only design vectors that would have led to a smaller reference statistic are: $(1,1,1,1,0,1,0,0,0,0)$ with a reference statistic of -27.4, $(1,1,1,0,1,1,0,0,0,0)$ with a reference statistic of -27, and $(1,1,0,1,1,1,0,0,0,0)$ with a reference statistic of -26.6. Because there are only three additional reference statistics smaller than or equal to the observed statistic, applying Equation (14) gives a *p*-value equal to 4/252 = .0159. In clinical trials, it usually cannot be ruled out that the treatment has a paradoxical effect, and therefore two-tailed tests are preferred. If the absolute value of the difference between the averages is taken as a test statistic, then $s_{obs} = s_1(\boldsymbol{x_1}, \boldsymbol{y}) = \left|\frac{1}{5}\boldsymbol{x'y} - \frac{1}{5}(\boldsymbol{1'} - \boldsymbol{x'})\boldsymbol{y}\right| = |68.2 - 94.4| = 26.2$. Applying Equation (12) gives a *p*-value of 8/252 = .0317. So the null hypothesis that there is no treatment effect can be rejected at the 5% significance level, because .0317 is not larger than .05 [Equation (15)]. There appears to be a statistically significant smaller average weight in the Treatment as compared to the Control. Figure 1 shows the reference distribution and indicates the area with the reference statistics that are larger than or equal to the observed statistic, in absolute value.

_____

Insert Figure 1 here

_____

Main features of the randomization test

The randomization test has several striking features, which set it apart from other inferential procedures: it is valid and powerful by construction, it can be used with any design and any test statistic, without random sampling and without assuming specific population distributions, and it results in a frequentist, conditional, and causal inference.

- ***Validity and power by construction***

The randomization test is exactly valid by construction. If the null hypothesis is true and virtual repetitions of the trial are contemplated, the response vector is invariant. So whatever design vector would be applied, the responses will always be identical if the null hypothesis is true. Because each design vector corresponds to one reference statistic, randomly selecting a design vector with probability 1/*K* implies randomly selecting a reference statistic with probability 1/*K*. Consequently, the probability of a particular subset of reference statistics can be calculated as the proportion of these reference statistics in the finite reference distribution, as is done in Equations (12) and (14).

If the decision rule of Equation (15) is used, then two kinds of errors can be made. An error of the first kind is made when the rule leads to a rejection of a true null hypothesis (a false alarm). An error of the second kind is made when the rule does not lead to a rejection of a false null hypothesis (a missed detection).

It can be shown that the randomization test controls the probability of an error of the first kind in the strict sense, which means that this probability is never larger than the level of significance α (Edgington & Onghena, 2007; Onghena, 1994). It is possible that the probability of an error of the first kind is substantially smaller than the level of significance α (i.e., it is a conservative test) because the reference distribution is discrete (with probability steps of 1/*K*) and because additional reference statistics may be tied with the observed statistic (Keller, 2012; Onghena 1994), but this conservatism can be made negligible by setting the level of significance α equal to a multiple of 1/*K* or by adding auxiliary randomization (Berger, 2000, 2004, 2007, 2008; Edgington & Onghena, 2007; Keller, 2012; Onghena, 1994). This error rate control holds for randomization tests in all randomized trials, even for trials that have very small *N* and even if $n_i = 1$ in some or all of the treatments. This feature distinguishes the randomization test from many other inferential procedures that have only guaranteed validity for large *N* or that need two or more observations in each treatment (because within-treatment variance has to be estimated).

Just like for any other inferential procedure, however, *N* is a crucial determinant of the probability of an error of the second kind for a randomization test. For an extreme example, if *N* is chosen such that $K < 20$ by Equation (1), then Equation (15) implies that a 5% level randomization test will result in an error of the second kind with absolute certainty (probability of 100%) because in that case it is impossible to have $p \leq \alpha$ for any true alternative hypothesis. Besides *N*, the statistical power of a randomization test (i.e., the complement of the probability of an error of the second kind) is function of the design, the vector of basic responses, the treatment effect, the test statistic, and the level of significance α (Gabriel & Hall, 1983; Gabriel & Hsu, 1983; Keller, 2012; Kempthorne & Doerfler, 1969). Because the most sensitive test statistic can be devised for the specific treatment effect that is expected, randomization tests with optimal power can be constructed in a variety of situations. For example in a completely randomized design with two or more than two treatments, if the alternative

hypothesis specifies random sampling from normally distributed populations with different population averages, the randomization test using *t* or *F* as a test statistic has an asymptotic relative efficiency of 100% as compared to the classical parametric Student's *t*- or *F*-test (Hoeffding, 1952). If a test statistic based on ranked transformed data is used (Wilcoxon-type statistics) then the asymptotic relative efficiency of the randomization test compared to Student's *t* test is never smaller than 0.864 and is larger for logistic, exponential, and double exponential distributions; for Cauchy distributions, the asymptotic relative efficiency is even infinitely large (Hollander, Wolfe, & Chicken, 2014; Lehmann & Romano, 2005).

- ***Flexibility with respect to the design and the test statistic***

A distinctive feature of the randomization test is that it can be used with any customized randomized design. Other statistical tests are usually linked to a particular conventional setting such as the one-sample problem, two independent samples, paired samples, the split-plot design and so on, but a randomization test can be applied to any design that fits the practical circumstances of the study under consideration, even if these circumstances call for an unconventional randomization procedure. For example, because of practical or ethical considerations it might be needed to randomly assign only one patient to a particular treatment, or the number of patients might be determined randomly, or the treatment-assignment probabilities might be unequal for some or all patients, or the treatment-assignment probabilities might be zero for particular treatment-patient combinations. Randomization tests can easily accommodate these circumstances by mimicking the randomization procedure in the derivation of the reference distribution (Edgington & Onghena, 2007). Randomization tests can even be applied to test for treatment effects in randomized *N*-of-1 clinical trials, which by definition employ unconventional randomization procedures (see Dugard, File, & Todman, 2012; Edgington, 1967; Ferron & Levin, 2014; Onghena, 1994, 2007, 2016; Onghena & Edgington, 2005).

Another distinctive feature of randomization tests is that they can be used with any customized test statistic. Other statistical tests have to use a specific test statistic, for example the *t*, the *F*, or the χ^2 statistic, for which the small-sample or asymptotic sampling distribution has been derived mathematically. For a randomization test, however, the reference distribution of any test statistic is derived specifically for each data set, by combining the response vector and the design vector. The test statistic is chosen freely to fit the purpose of the research and this makes the randomization test a very versatile inferential technique for a multiplicity of research contexts (Berger, 2000; Dugard, 2014). Because of this flexibility and versatility, the randomization test can be considered a generic procedure to derive a reference distribution rather than a narrowly defined statistical test. For example, a *t* test statistic can be used in a randomization test (to obtain a so-called "randomization *t* test"), but we can also use the difference between medians or the variance ratio as test statistics if those statistics more closely fit the researcher's hypothesis (Edgington & Onghena, 2007).

- ***No random sampling and no distributional assumptions required***

The most prominent feature of a randomization test is that no random sampling is assumed. Consequently, in research situations in which random sampling is impractical or impossible, a randomization test is best suited (Edgington, 1966; Hunter & May, 2003; Ludbrook & Dudley, 1998). For example, in clinical trials it is usually not realistic to assume that the patients have been sampled from a hypothetical infinitely large population; patients enter the trial one-by-one based on hospital

admissions and trial selection criteria, and constitute a convenience sample rather than a random sample (Berger, 2000; Cleophas, Zwinderman, Cleophas, & Cleophas, 2009; Onghena & Edgington, 2005).

Furthermore, because no random sampling is assumed, there is no need for a population model, and consequently there is also no need for assumptions about population distributions. The randomization test is defined within a so-called "randomization model" that only needs the observed data, a randomized design, and a sensitive test statistic (Keller, 2012, Kempthorne, 1955; Lehmann, 1975; Lehmann & Romano, 2005; Pitman, 1937a, 1937b, 1938). In the same vein, notice that the null hypothesis in Equation (7) is not formulated in terms of population parameters, but that it is formulated at the level of the observed responses of every single patient (Edgington & Onghena, 2007).

- ***Frequentist, conditional, and causal inference***

A final characteristic of the randomization test concerns the nature of the statistical inference that is achieved. First of all, this inference is frequentist (Bandyopadhyay & Forster, 2011; Wasserman, 2008; Zelen, 1998). The randomization test fits within the frequentist inferential frameworks of Fisher and Neyman-Pearson. The *p*-value in Equation (12) or Equation (14) can be interpreted directly in frequentist terms: This *p*-value is the relative frequency of reference statistics that are as extreme as, or more extreme than, the observed statistic in the reference distribution. In other words, it is the relative frequency of trials that are at the same distance or even more distant to the null hypothesis than the observed trial if an infinite number of repetitions of the trial for a true null hypothesis is conceived of. Furthermore, the operating characteristics of the randomization test can be expressed as long-term first and second kind error rates (Hinkelmann & Kempthorne, 2005, 2008, 2012; Lehmann, 1975).

Secondly, the inference from a randomization test is conditional on the observed data, just like Fisher's Exact Test is conditional on the marginal totals (Hothorn, Hornik, van de Wiel, & Zeileis, 2006; Krauth, 1988; Pesarin & Salmaso, 2010). In fact in a randomized trial, Fisher's Exact Test is a randomization test for comparing two treatments with a dichotomous outcome variable (Agresti, 2013; Edgington & Onghena, 2007; Mehta, 1994). This conditionality is obvious from the definition of the test statistic, $s_k = s(\boldsymbol{x_k}, \boldsymbol{y})$, as a function s of a random design vector and a fixed response vector (no subscript k for vector $\boldsymbol{y}$).

Lastly and most importantly, the inference from a randomization test is causal. The null hypothesis of no treatment effect can be interpreted as a null hypothesis about the absence of a causal relation. The randomization test itself focuses on the functional relation between the manipulated factor *X* and the outcome variable *Y*, and derives a *p*-value by counterfactual reasoning (Holland, 2005; Lewis, 1973a, 1973b). Suppose that *X* has only two levels (Treatment T and Control C) then the causal effect of *X* on outcome *Y* in a clinical trial is defined as the difference in *Y* between T and C for a given patient. The average causal effect is the average difference over all patients. The fundamental problem of causal inference is that it is impossible in a between-subjects group comparison study to observe a patient in T and C simultaneously. One of the outcomes is observed, and the other one is missing. The randomization test offers an elegant solution to this fundamental problem: If the null hypothesis is true, we know both outcomes because the outcome *Y* in T is identical to the outcome *Y* in C (Holland, 1986; Imbens & Rubin, 2015; Rubin, 1974).

TWO SIDES OF THE SAME COIN

After this presentation of the randomization test, we can more fully appreciate the link between randomization and the randomization test. This link is obvious when looking closely at Fisher's famous quote about the conclusion we can draw from a significance test:

> The force with which such a conclusion is supported is logically that of a simple disjunction: Either an exceptionally rare chance has occurred, or the theory of random distribution is not true. (Fisher 1956, p. 39)

In this quote we can interpret the "conclusion" as the "rejection of the null hypothesis" and "the theory of random distribution" as "the null hypothesis". So Fisher (1956) explicitly conceptualized the conclusion drawn from a significance test as an exclusive disjunction: if we obtain a small *p*-value then either we witnessed an improbable event or the null hypothesis is not true. The improbable event happens in a clinical trial if the patients are assigned in such a way that "by accident" more favorable basic values end up in the treatment condition and more unfavorable ones end up in the control condition, erroneously suggesting a treatment effect that is nonexistent. In the hypothetical example of testing a new diet against a control, this would mean that the new diet has no effect but that "by accident" patients that have less weight have been assigned to the treatment condition. In Table 1, this happens in Assignment 1, with patients weighing 62 kg and 75 kg assigned to Treatment and patients weighing 76kg and 127 kg assigned to Control. If these weights were the observed responses in a randomization test, then the one-tailed *p*-value would be $1/K = .125$. In virtual repetitions of this trial, this event is expected to happen 12.5% of the time. By setting the level of significance α at a small amount we can control the first kind error rate at a small and known value.

In sum, the randomization test derives its statistical validity by virtue of the randomization as actually used in the clinical trial. Conversely, if a randomized design is used, it makes sense to complement the trial and measurement of the outcomes with a calculation of the randomization test *p*-value because it provides a quantification of the probability of these outcomes if the null hypothesis is true. In this way, randomization and the randomization test are intimately linked: they are two sides of the same coin.

REFERENCES

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Bandyopadhyay, P. S., & Forster, M. R. (Eds.) (2011). *Handbook of the philosophy of science, Vol. 7: Philosophy of statistics.* Amsterdam, The Netherlands: Elsevier.

Basu, D. (1975). Statistical information and likelihood [with Discussion]. *Sankhyā: The Indian Journal of Statistics, Series A, 37,* 1–71.

Basu, D. (1980). Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association, 75,* 575–582.

Begg, C. B. (2015). Ethical concerns about adaptive randomization. *Clinical Trials, 12,* 101.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schulz, K. F., Simel, D., & Stroup, D. F. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association, 276,* 637–639.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine, 19,* 1319–1328.

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials, 25,* 613–619.

Berger, V. W. (2007). Drawbacks to non-integer scoring for ordered categorical data. *Biometrics, 63,* 298–299.

Berger, V. W. (2008). Letter to the editor. *Biometrical Journal, 50,* 1–3.

Berger, V. W., & Bears, J. D. (2003). When can a clinical trial be called 'randomized'? *Vaccine, 21,* 468–472.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2014). *A chronicle of permutation statistical methods: 1920–2000, and beyond.* New York, NY: Springer.

Berry, S., Carlin, B., Lee, J., & Müller, P. (2010). *Bayesian adaptive methods for clinical trials.* Boca Raton, FL: CRC Press.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Cleophas, T. J., Zwinderman, A. H., Cleophas, T. F., & Cleophas, E. P. (2009). Clinical trials do not use random samples anymore. In Cleophas, T. J., Zwinderman, A. H., Cleophas, T. F., & Cleophas, E. P. (Eds.), *Statistics applied to clinical trials* (4th ed.) (pp. 367–374). New York, NY: Springer.

Cochran, W., & Cox, G. (1950). *Experimental designs.* New York, NY: Wiley.

Cochrane, A. L. (1972). *Effectiveness and efficiency: Random reflections on health services.* London, UK: Nuffield Provincial Hospitals Trust.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago, IL: Rand McNally.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics.* London, UK: Chapman and Hall.

Cox, D. R., & Reid, N. (2000). *The theory of the design of experiments.* Boca Raton, FL: Chapman & Hall/CRC.

David, H. A. (1995). First (?) occurrences of common terms in mathematical statistics. *The American Statistician, 49,* 121–133.

David, H. A. (2008). The beginnings of randomization tests. *The American Statistician, 62,* 70-72.

Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science, 3,* 65–68.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests* (2nd ed.). New York, NY: Routledge Academic.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin, 66,* 485–487.

Edgington, E. S. (1967). Statistical inference from *N*=1 experiments. *The Journal of Psychology, 65,* 195–199.

Edgington, E. S. (1970). Hypothesis testing without fixed levels of significance. *The Journal of Psychology, 76,* 109–115.

Edgington, E. S. (1986). Randomization tests. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 7* (pp. 530–538). New York, NY: Wiley.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Ferron, J. M., & Foster-Johnson, L. (1998). Analyzing single-case data with visually guided randomization tests. *Behavior Research Methods, Instruments, & Computers, 30,* 698–706.

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple baseline data. *The Journal of Experimental Education, 75,* 66–81.

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Statistical and methodological advances* (pp. 153–183). Washington, DC: American Psychological Association.

Finch, P. D. (1986). Randomization-I. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 7* (pp. 516–519). New York: Wiley.

Fisher, R. A. (1925). *Statistical methods for research workers.* London, UK: Oliver & Boyd.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33,* 503–513.

Fisher, R. A. (1935). *The design of experiments.* London, UK: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference.* London, UK: Oliver & Boyd.

Folks, J. L. (1984). Use of randomization in experimental research. In K. Hinkelmann (Ed.), *Experimental design, statistical models, and genetic statistics: Essays in honor of Oscar Kempthorne* (pp. 17–32). New York, NY: Marcel Dekker.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine, 317,* 141–145.

Gabriel, K. R., & Hall, W. J. (1983). Rerandomization inference on regression and shift effects: Computationally feasible methods. *Journal of the American Statistical Association, 78,* 827–836.

Gabriel, K. R., & Hsu, C.-F. (1983). Evaluation of the power of rerandomization tests, with application to weather modification experiments. *Journal of the American Statistical Association, 78,* 766–775.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York, NY: Springer.

Greenberg, B. G. (1951). Why randomize? *Biometrics, 7,* 309–322.

Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis, 79,* 427–451.

Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions: Version 5.1.0* [updated March 2011]. The Cochrane Collaboration. Available from http://handbook.cochrane.org.

Hinkelmann, K., & Kempthorne, O. (2005). *Design and analysis of experiments, Vol. 2: Advanced experimental design*. Hoboken, NJ: Wiley.

Hinkelmann, K., & Kempthorne, O. (2008). *Design and analysis of experiments, Vol. 1: Introduction to experimental design* (2nd ed.). Hoboken, NJ: Wiley.

Hinkelmann, K., & Kempthorne, O. (2012). *Design and analysis of experiments, Vol. 3: Special designs and applications.* Hoboken, NJ: Wiley.

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. *Annals of Mathematical Statistics, 23,* 169–192.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

Holland, P. W. (2005). Counterfactual reasoning. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science, Vol. 1* (pp. 420–422). Chichester, UK: Wiley.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric statistical methods* (3rd ed.). Hoboken, NJ: Wiley.

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2006). A Lego system for conditional inference. *The American Statistician, 60,* 257–263.

Hu, F., & Rosenberger, W. (2006). *The theory of response-adaptive randomization in clinical trials.* Hoboken NJ: Wiley-Interscience.

Hunter, M. A., & May, R. B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 57,* 176–188.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences: An introduction.* New York, NY: Cambridge University Press.

Jadad, A. R., & Enkin, M. W. (2007). *Randomized controlled trials: Questions, answers and musings* (2nd ed.). Oxford, UK: Blackwell.

Kadane, J. B., & Seidenfeld, T. (1990). Randomization in a Bayesian perspective. *Journal of Statistical Planning and Inference, 25,* 329–345.

Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika, 77,* 324–338.

Kempthorne, O. (1952). *The design and analysis of experiments.* New York, NY: Wiley.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association, 50,* 946–967.

Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference, 1,* 1–25.

Kempthorne, O. (1986). Randomization-II. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 7* (pp. 519–524). New York: Wiley.

Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika, 56,* 231–248.

Kempthorne, O., & Folks, J. L. (1971). *Probability, statistics, and data analysis.* Ames, IA: Iowa State University Press.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* San Francisco, CA: Holden-Day.

Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach.* New York, NY: Elsevier.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). New York, NY: Springer.

Lewis, D. (1973a). Causation. *Journal of Philosophy, 70,* 556–567.

Lewis, D. (1973b). *Counterfactuals.* Oxford: Blackwell.

Lindley, D. V. (1982). The role of randomization in inference. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, 2,* 431–446.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* Tests in biomedical research. *The American Statistician, 52,* 127–132.

Manly, B. F. J. (2007). *Randomization, bootstrap and Monte Carlo methods in biology* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Matthews, J. R. (this volume). *Randomization and bias in historical perspective*.

Mehta, C. R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research, 3,* 135–156.

Onghena, P. (1994). *The power of randomization tests for single-case designs.* Unpublished doctoral dissertation, Department of Psychology, Katholieke Universiteit Leuven, Belgium.

Onghena, P. (2007). *N-of-1 randomized clinical trials.* In The Biomedical & Life Sciences Collection, Henry Stewart Talks Ltd, London (online at https://hstalks.com/t/555/)

Onghena, P. (2016). *Randomization in N-of-1 clinical trials: Is it possible to draw causal inferences from single-patient data?* In The Biomedical & Life Sciences Collection, Henry Stewart Talks Ltd, London (online at https://hstalks.com/bs/3311/)

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56–68.

Peirce, C. S., & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences, 3,* 73–83.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software.* Chichester, UK: Wiley.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society Series B, 4,* 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations II: The correlation coefficient. *Journal of the Royal Statistical Society Series B, 4,* 225–232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations III: The analysis of variance test. *Biometrika, 29,* 322–335.

Rosenberger, W. F., & Lachin, J. M. (2015). *Randomization in clinical trials: Theory and practice* (2nd ed.). Hoboken, NJ: Wiley.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics, 6,* 34–58.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics, 2,* 808–840.

Saint-Mont, U. (2015) Randomization does not help much, comparability does. *PLoS ONE, 10*(7). e0132102. doi:10.1371/journal.pone.0132102

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York, NY: W. H. Freeman.

Savage, L. J. (1962). Subjective probability and statistical practice. In G. A. Barnard & D. R. Cox (Eds.), *The foundations of statistical inference: A discussion* (pp. 9–35). London, UK: Methuen.

Senn, S. (1994). Fisher's game with the devil. *Statistics in Medicine, 13,* 217–230.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Stigler, S. M. (1978). Mathematical statistics in the early states. *Annals of Statistics, 6,* 239–265.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Stone, M. (1969). The role of experimental randomization in Bayesian statistics: Finite sampling and two Bayesians. *Biometrika, 56,* 681–683.

Wasserman, L. (2008). Comment on article by Gelman. *Bayesian Analysis, 3,* 463–466.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70,* 129–133.

Zelen, M. (1998). Inference. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 2035–2046). New York, NY: Wiley.

Table 1

Hypothetical Example of the Equating Property of Randomization in a Completely Randomized Design with Two Treatments and Two Patients in Each Treatment

	Treatment		Control		Average
Assignment 1	62	75	76	127	340/4 = 85
Assignment 2	62	76	75	127	340/4 = 85
Assignment 3	62	127	75	76	340/4 = 85
Assignment 4	75	76	62	127	340/4 = 85
Assignment 5	75	127	62	76	340/4 = 85
Assignment 6	76	127	62	75	340/4 = 85
Average	1020/12 = 85		1020/12 = 85		2040/24 = 85

Table 2

Hypothetical Example of Table 1 with Averages for Each Treatment and the Deviation for Each Assignment

	Treatment	Control	Deviation
Assignment 1	68.5	101.5	−33
Assignment 2	69	101	−32
Assignment 3	94.5	75.5	19
Assignment 4	75.5	94.5	−19
Assignment 5	101	69	32
Assignment 6	101.5	68.5	33
Average	510/6 = 85	510/6 = 85	0/6 = 0

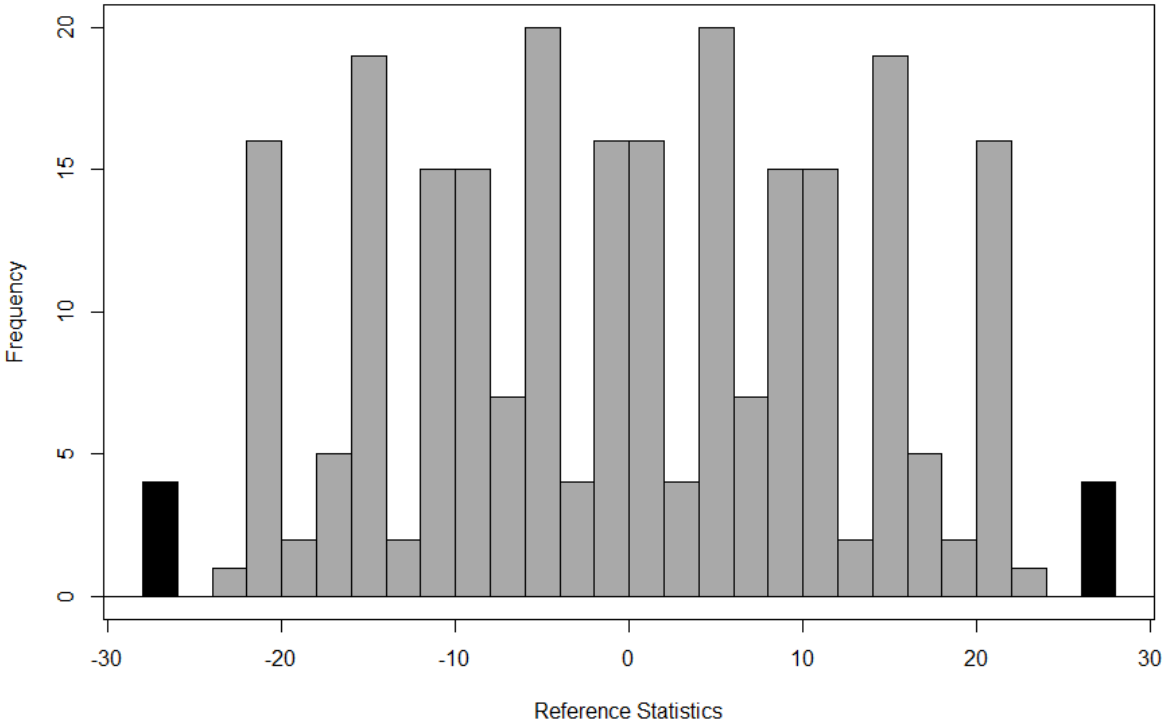

Figure 1. Reference distribution for a randomization test in a completely randomized design with two treatments and five patients in each treatment. The data are 51, 62, 75, 76, and 77 for one treatment and 74, 89, 90, 92, 127 for the other, and the difference between the treatment averages is taken as the test statistic. The black bars correspond to the reference statistics that are larger than or equal to the observed statistic, in absolute value. The two-tailed randomization test *p*-value corresponds to 8/252.